IMPLEMENTATION OF ORD S PRESERVATION, ARCHIVE AND RELEASE POLICY

A STRATEGIC WHITE PAPER

Linda Kirkland
NCERQA/QAD
December, 1997

EXECUTIVE SUMMARY

The EPA Strategic Plan (EPA,1997) recognizes that science enables us to identify the most important sources of risk to human health and the environment, and by so doing, informs our priority-setting, ensures credibility for our policies, and guides our deployment of resources . This emphasizes the importance of scientific research to EPA s function. One of the chief characteristics of science is reproducibility of research results. ORD recognizes this by proposing to adopt a policy for intramural (and possibly extramural) research that records and research materials must be retained in sufficient detail so that individuals trained in the appropriate disciplines can reconstruct the research (Cortesi, November 6, 1997 draft). The proposed policy also requires that data should have been through QA procedures .

This policy will need considerable effort in planning and coordination of data management activities both among ORD organizations and with other external organizations to be implemented. The ORD Strategic Plan s Information Management Component (ORD, September, 1997) recognizes that ORD s success is defined by two factors 1) the scientific quality of our research & development and 2) the degree to which EPA and external stakeholders have open access to and can use the information and data

we generate. In the document, science quality is addressed through the interface with EPA s mandatory Quality Assurance (QA) program [ORD usually includes peer review] which generates records on research organizations quality systems, research QA project plans, and documentation that the environmental measurements funded by EPA are of the level of quality suitable for their intended use such as drawing research conclusions. To be usable, data resulting from environmental measurements need to have associated information (scientific meta-data) that defines what the data sets represent and how well they do it that is accessible to potential users so they can make informed judgements on whether or not to use the data.

Several issues impact ORD s capability to document research sufficiently to provide reproducibility:

1. ORD s record retention schedules relate to Federal Records Center (FRC) archived paper files and do not specify a mandatory level of detail in documentation for pedigrees of data of on-going research use or temporal linkage of program, data files, system and application documentation and QA records for storage at alternative sites like the National Computer Center.

2. EPA has not followed up the proposal by EPA s Office of Information Resources Management (OIRM) (Johnson, 1993) to adopt a scientific meta-data policy.

3. Research data are not contained in searchable databases to provide ready access for uses supporting ORD research strategies.

4. Definition in data dictionaries and registry of data elements for environmental measurements and their associated quality control (QC) information is incomplete and could prevent access and use by ORD and its customers.

A strategy is presented to address these issues and implement ORD s preservation, archive and release policy through the Science Information Management Coordination Board (SIMCorB).

PROBLEM DEFINITION
BACKGROUND

EPA is addressing changes in Federal regulations on information management. ORD, for its part, is taking greater management responsibility in this area to assure compliance with the Information Technology Management Reform Act of 1996 by developing its Strategic Plan s IM Component and establishing the SIMCorB board to implement it. Data Administration and Quality Assurance (DA/QA) have been incorporated as a board sub-group in SIMCorB s charter. Coordination of ORD policy development in areas of scientific meta-data and QA activities supporting data usability and integrity, respectively, are part of the sub-group s mission. Development of an infrastructure to provide needed standards and uniform policies across ORD and feedback to ORD

management on the success of their implementation is a goal of the DA/QA sub-group.

The IM Component of the ORD Strategic Plan (http:/www.epa.gov/ORD) notes under Strategy Component 4 that for utility to be optimized ORD must ensure that adequate documentation can be obtained so that ORD data and information that is located and retrieved is also usable. One area needing to be addressed is scientific meta-data or the information describing the purpose for which the data were collected and the pedigree of how they were produced and assessed for adequacy for that use (i.e., quality). This involves keeping adequate records of research planning and documentation of assessments of plan implementation and data quality as well as providing access to such records.

QA PROGRAM

EPA s mandatory QA program provides requirements consistent with the American National Standard, Specifications and Guidelines for Quality Systems for Environmental Data Collection and Environmental Technology Programs (ANSI/ASQC E4-1994). For example, the EPA Requirements for QA Project Plans for Environmental Data Operations (EPA QA/R-5, Quality Assurance Division (QAD), October 1997) include the following data management components in addition to requirements for data quality objectives, methods for data collection and analysis, as well as assessments of plan implementation and data quality:

- **Describe the project data management scheme, tracing the path of the data from their generation in the field or laboratory to their final use or storage.**
- **Describe or reference the standard record-keeping procedures, document control system, and the approach used for data storage and retrieval on electronic media.**
- **Discuss the control mechanism for detecting and correcting errors and for preventing loss of data during data reduction (i.e., calculations), data reporting, and data entry to forms, reports, and databases.**
- **Identify and describe all data handling equipment and procedures to process, compile, and analyze data. This includes procedures for addressing data generated as part of the project as well as data from other sources.**
- **Include any required computer hardware and software and address any specific performance requirements for the hardware/software configuration used. Describe the procedures that will be followed to demonstrate acceptability of the hardware/software configuration required.**
- **Describe the process for assuring that applicable Agency information resource management requirements are satisfied (e.g., EPA Directive 2100 and EPA Orders 2180, 2180.1, 2180.2, & 7500.1A).**

These required QA project plans as part of research planning and the documentation of their implementation (e.g., assessment or test reports and QC data) provide critical records of research implementation and resulting data quality.

INTERFACE WITH OIRM S AGENCY RECORD MANAGEMENT, SCIENTIFIC META-DATA POLICY RECOMMENDATION AND DATA ELEMENT REGISTRY

The EPA Records Officer leads Agency records planning but relies on the specific programs like ORD to assist in development of policy for electronic records such as scientific databases (e.g., determining retention period and implementing authorized disposition instructions) because current policy specifies the responsibility for creation and use of systems records to information systems (program) managers (Information Resources Management Policy Manual, Directive 2100, Records Management Chapter 10, July 1996 see http:/www.epa.gov/nrmp). ORD has not developed schedules for all of its database systems (e.g., IM Component list) or applications (e.g., Schedule 063A) and many that exist are not approved by NARA. Even those that are approved (e.g., 466A) only specify what database records it can include (e.g., software source code, data system specifications, file specifications, user guides and output specifications) and store on tape once systems and programs are superceded. It would be impossible to reconstruct summary data sets without detailed information on what the raw data was and how it was processed.

Specifications for records adequate to provide a complete pedigree for a research database are currently lacking. QA records such as QA project plans (185A) and ORD Laboratory Performance Evaluation Studies (586L) have not been approved (10/14/97 list). Other records like data quality and technical systems audit reports and sampling or analysis QC data have not been scheduled. Also some schedules need to be revised to be linked and temporally consistent. For example, data files of continuing research interest are to be retained for 20 years after their transfer to FRC (503L), or permanently (469A); lab notebooks are scheduled to be destroyed five years after project completion; approved QA project plans are proposed to be kept 10 years; software programs can be destroyed when updates and their QA checks are completed (457H) or after 7 years (462A). Specific database schedules following NDPD s policy manual state that supporting documentation can include records such as software source code with storage on tape for 7 years (466A).

The Government Information Locator System guidelines for meta-data provide only general examples under the heading Methods that would describe how scientific data sets were derived. OIRM has not followed-up on its recommended actions for improving the usefulness of the Agency s electronically stored data by developing policy on required scientific meta-data content and access (Meta-data Recommendations Report Memorandum, June 14, 1993). ORD should review this policy proposal (e.g., storage of data quality indicators and pedigrees with scientific data) in light of the need to have records sufficient to reconstruct its research and assure usability of its data. The major issues for scientific meta-data policy development involve imposing requirements for meta-data to be entered along with scientific data in data sets of continued research interest (e.g., cost to data collectors) and identifying what meta-data elements are needed to support data quality assessments of both primary and secondary users (e.g., benefit to others). Some data collectors are not under the jurisdiction of ORD (e.g., state and Regional programs and other Federal Agencies) and data ownership concerns arise when ORD does not or only partially funds the project.

The EPA Office of Information Resources Management s Environmental Information Management Division (EIMD) is beginning to address entering scientific meta-data elements into the EPA data element registry (http:/www.epa.gov/edt). Defining such elements so that they can be standardized across EPA databases facilitates sharing and integration of data (e.g., across media). EPA s Ecological Research Strategy (ORD, June 1997) and the Human Health Risk Research Strategy (ORD, June 1997) drafts both emphasize the need to use and integrate data collected by others in the development of methods such as indicators and models. They will need to characterize and assess such data for adequacy for their intended uses such as data aggregation and modeling. Therefore, secondary data users in ORD need to be involved in this effort which would help identify and define needed scientific meta-data elements.

PROPOSED APPROACH TO PROBLEM SOLUTION

The DA/QA subgroup offers an opportunity for strategic coordinating within SIMCorB and with OIRM and the ORD QA managers to advance efforts to develop standards and policies supporting sufficient record keeping to allow reconstruction of its research and access to such information as scientific meta-data. SIMCorB represents all ORD organizations in advising the ORD Assistant Administrator and addresses all stages of the life cycle of database development in its sub-groups. The DA/QA sub-group addresses scientific meta-data as well as software/hardware configuration QA, security, and configuration management (e.g., version control of systems and software applications like models). The chair of the DA/QA sub-group, a staff member of ORD s QAD, has experience in data management QA policy development and implementation for the EPA s QA program in general and specifically within ORD.

The first phase of solution would involve enlisting experienced scientists and system developers in a joint application design effort to identify useful scientific meta-data elements and feasible system design for their access and use in data quality characterization and assessment. Selected scientists would be instructed on QAD s data quality assessment guidance and target example assessments (e.g., use of secondary data in indicator or model development). These examples would be used in identification of useful records or meta-data elements to support processing functions in data quality assessments. QAD would provide training and assist in initial planning for this activity in coordination with the SIMCorB DA/QA and Requirements sub-groups.

SIMCorB has a pilot, the Environmental Information Management System (EIMS of NCEA and Region 10) that has developed a Directory/Catalog/Dictionary approach as a prototype design for the management of descriptive information about data, projects, and models developed or held by EPA (e.g., scientific meta-data entry and access). This pilot is coordinated with the Regional Vulnerability Assessment (ReVA) database where indicators are being developed in the Mid-Atlantic Integrated Assessment project. The latter project is a prototype for multi-media, multi-scale and integrated compartment models needing advanced data access, and data management, as well as modeling and model output interpretation application capabilities for a specific geographic area. A

pertinent example could be developed within the prototype to present to the selected scientists and system developers to encourage requirements development. The benefits of access to this information in meeting ORD research records policy for reconstructing research and use in methods development would be identified for marketing the development efforts to ORD and non-ORD Agency management.

Once potential data elements are identified, definition could be coordinated with EIMD s registry group and the existing or draft record retention schedules for electronic data consistent with EPA s developing policy through the EPA records officer. System development within the EIMS/ReVA pilot would be coordinated with the Systems Engineering/Operations and Advanced Technology sub-groups. DA/QA would pursue standards and policies for ORD to 1) assure entry of adequate scientific meta-data with data sets from required records (e.g., QA plans and assessments) and 2) the maintenance of their integrity through adequate controls of system and application development, configuration management and security. These efforts would be coordinated with the Engineering/Operations sub-group in proposing roles and responsibilities for ORD personnel such as scientific researcher users, extramural project managers, data administrators, ADP coordinators and security officers as well as QA managers. Oversight of implementation of these standards and policies would be coordinated with the Engineering/Operations sub-group (e.g., ADP coordinators) and ORD QA managers. For example, revision of the ORD Management Information System (OMIS) database for inclusion of QA activity tracking starting with additions to the NRMRL QA pilot module of system and application development requirement documents and corresponding test results.